摘要

由于帧相机成像机制的限制，基于常见的彩色相机或者深度相机的手势姿态估计在强光和快速运动场景下是不可靠的。作为一种新兴的神经形态传感器，事件相机由于具有高动态范围、高时间分辨率、低功耗的特性，在手势姿态估计领域富有潜力。但是事件相机难以捕获纹理信息，因此在静止或者光照变化的场景存在短板。事件相机和帧相机的这些优势和缺点启发本工作将两种模态的数据进行融合，利用各自的优势弥补对应的短板，实现稳定、高精度且高效的手势姿态估计。

本文研究了如何使用事件相机和帧相机实现鲁棒的手势姿态估计，完成了两项工作：基于事件流的手势姿态估计，基于事件流与 RGB 帧互补的手势姿态估计。

1. 本文提出了基于事件流的单目手势姿态估计方法。为了处理事件流的异步格式以及运动歧义问题，本文设计了针对手的光流表示刻画在事件相机观测下手的运动过程。相对于高时间分辨率的事件流，现有方法获得的真值标注是稀疏的，为充分利用事件流在时间维度上丰富的信息，本文设计了基于对比度最大化和边缘信息约束的自监督框架。为了验证提出的方法，本文采集了包含 74 分钟事件序列和 42.1 万张 RGB 帧的首个带有高精度标注的真实数据集。在真实数据集上的实验结果表明，本文提出的方法在快速运动场景和强光场景下表现超过现有的基于 RGB 帧的方法，而且可以实现 120 FPS 的手势姿态估计，定量与定性地证明了事件相机在手势姿态估计领域的潜力。
2. 本文提出了首个基于事件流和 RGB 帧互补的手势姿态估计方法。基于上个工作对 RGB 帧和事件流的各自优势场景以及相应挑战性问题的研究，本文设计了场景关联融合模块引入该先验知识，用事件流提升基于 RGB 帧的手势姿态估计在过曝光和运动模糊问题下的表现；考虑到事件流与 RGB 帧数据模态上差异，在绝大部分时刻事件流没有与之同步的 RGB 帧，本文设计了非同步融合模块将之前最近的 RGB 信息融合到当前事件流中，提升基于事件流的手势姿态估计在前景稀疏和背景溢出问题下的表现。实验表明，本文提出的方法可以通过两种数据的互补有效提升在挑战性场景下的效果。由于本方法是基于卷积神经网络和 Transformer 的结构，通过调整网络结构可以灵活实现精度和计算成本的权衡。

综上所述，本文对基于事件流和 RGB 帧的手势姿态估计进行了探索，从相机成像机制出发挖掘两种数据互补的潜力。本文证明了事件相机与帧相机在手势姿态估计任务上互补的技术可行性，为设计高效鲁棒的手势姿态估计系统提供了参考。

关键词：计算机视觉，手势姿态估计，事件相机

Complementing Event Streams and RGB Frames for Hand Pose Estimation

Jianping Jiang (Computer Science and Technology (Intelligence Science and Technology))
Directed by: Assistant Prof. Boxin Shi

ABSTRACT

Due to the limitations of frame-based camera imaging mechanism, hand pose estimation based on common-used color or depth sensors is unreliable in strong light and fast motion scenarios. As an emerging neuromorphic sensor, event cameras are promising in the field of hand pose estimation due to their high dynamic range, high temporal resolution, and low power consumption. However, event cameras have difficulty in capturing texture information and thus have a shortcoming in stationary or lighting change scenes. These advantages and disadvantages of event cameras and frame-based cameras inspire us to fuse data from both modalities and use their respective advantages to compensate for the corresponding issues to achieve stable, accurate and efficient hand pose estimation.

This thesis investigates how to achieve robust hand pose estimation using both event and frame-based cameras, and accomplish two works: event-based hand pose estimation, and hand pose estimation based on the complementary event stream and RGB frames.

1. This thesis proposes an event-based method for monocular hand pose estimation. To deal with the asynchronous format of event streams and the motion ambiguity issue, this thesis designs novel hand flow representations specially for the hand to portray the hand motion. The 3D annotations obtained by existing methods are sparse relative to the event streams with high temporal resolution. To make full use of the rich temporal information of the event streams, this thesis designs a self-supervised learning framework based on contrast maximization and edge constraints. To validate the proposed method, the first real-world dataset with accurate annotation containing 74 minutes of event sequences and 421 K RGB frames is collected in this work. Experimental results on the real-world dataset show that the proposed method outperforms existing RGB-based methods in fast motion scenes and strong light scenes, and can achieve 120 FPS hand pose estimation, which quantitatively and qualitatively demonstrate the potential

of event cameras in the field of hand pose estimation.

2. This thesis further proposes the first hand pose estimation method based on complementary event streams and RGB frames. Based on the previous work of RGB frames and event streams on the respective merits and challenging issues, this thesis designs a scene-aware fusion module to introduce this prior knowledge and uses event streams to improve the performance of RGB-based hand pose estimation under overexposure and motion blur issues. Considering the difference in data modality between event streams and RGB frames, the event streams do not have synchronized RGB frames. This thesis designs an non-concurrent fusion module to fuse the previous RGB information into the current event stream to improve the performance of event-based hand pose estimation under the foreground sparsity and background overflow issues. Experiments show that the proposed method can effectively improve the performance in challenging issues by complementing the two modality data. Since the method is based on the structure of convolutional neural network and Transformer, the trade-off between accuracy and computational cost can be flexibly achieved by adjusting the network structure.

In summary, this thesis explores hand pose estimation based on event streams and RGB frames, and demonstrate the complementary potential in terms of camera imaging mechanisms. This thesis shows the technical feasibility of complementing event and frame-based cameras for hand pose estimation tasks, and provides a reference for designing efficient and robust hand pose estimation systems.

KEY WORDS: Computer Vision, Hand Pose Estimation, Hand Mesh Reconstruction, Event Camera